Thurstonian Scaling of Compositional Questionnaire Data

Anna Brown

University of Kent

Author Note

Anna Brown, PhD, Senior Lecturer in Psychological Methods and Statistics, School of Psychology, University of Kent.

Correspondence should be addressed to Anna Brown, School of Psychology, University of Kent, Canterbury, Kent CT2 7NP, United Kingdom. E-mail: A.A.Brown@kent.ac.uk

Abstract

To prevent response biases, personality questionnaires may use comparative response formats. These include forced choice, where respondents choose among a number of items, and quantitative comparisons, where respondents indicate the extent to which items are preferred to each other. The present article extends Thurstonian modeling of binary choice data (Brown & Maydeu-Olivares, 2011a) to "proportion-of-total" (compositional) formats. Following Aitchison (1982), compositional item data are transformed into log-ratios, conceptualized as differences of latent item utilities. The mean and covariance structure of the log-ratios is modelled using Confirmatory Factor Analysis (CFA), where the item utilities are first-order factors, and personal attributes measured by a questionnaire are second-order factors. A simulation study with two sample sizes, N=300 and N=1000, shows that the method provides very good recovery of true parameters and near-nominal rejection rates. The approach is illustrated with empirical data from N=317 students, comparing model parameters obtained with compositional and Likert scale versions of a Big Five measure. The results show that the proposed model successfully captures the latent structures and person scores on the measured traits.

Keywords: Thurstonian factor models, compositional data, multiplicative ipsative data

Thurstonian Scaling of Compositional Questionnaire Data

In personality and similar assessments that rely heavily on respondent-reported measures, comparative judgments may be preferred to absolute judgments. This is because direct comparisons between questionnaire items facilitate differentiation and calibration (Kahnemann, 2011), thus reducing halo effects, and remove uniform biases such as acquiescence and leniency (Cheung & Chan, 2002). The most popular comparative format – *forced choice* – requires participants to select one of two items, or rank three or more items within so-called *blocks*. Until recently, forced-choice items have been scored by considering their relative positions in blocks – and thus yielding *ipsative* data, which is characterized by the total score on the test being the same for everyone (e.g., Clemans, 1966). Ipsative scores are centered on the persons' mean, and obviously present a problem in applications where inter-individual comparisons are sought. Thurstonian IRT models (Brown & Maydeu-Olivares, 2011a) were developed to overcome the problems of ipsative data in multidimensional forced-choice questionnaires. The approach uses SEM with categorical outcomes to model the mean and covariance structure of pairwise decisions – binary observed variables reflecting choices (prefer A to B, or prefer B to A) in every pair of items that respondents compare within forced-choice blocks. The pairwise decisions in these models are underlain by latent item *utilities* being compared (Thurstone, 1927; 1929), which in turn are underlain by psychological attributes the items are designed to measure.

Simple choice, however, is not the only way of expressing preferences between items. Quantitative information about the extent of preferences can also be captured. In *compositional* preference tasks, respondents have to distribute a fixed number of points (for instance, 100) between several items according to the extent the items describe their personality or represent their attitude, etc. A questionnaire may be compiled of many such tasks (compositions). With this format, preferences are expressed as proportions with respect

to a common basis, also dubbed *multiplicative* ipsative data (Chan, 2003). This type of ipsative data requires a model that would capture the quantitative information contained in the compositions, and would enable proper scaling of psychological attributes to allow inter-individual comparisons. The aim of the present article is to develop such a model.

Attempts to analyze the general covariance structure of multiplicative ipsative data in psychometric applications have been made before. Chan and Bentler (1993) used the first-order Taylor series approximation to restore the "true" pre-ipsative covariance structure of a single compositional task, which did not yield results of desired accuracy. More recently, Coenders, Hlebec and Kogovsek (2011) applied a general statistical approach of Aitchison (1982), mostly known outside of psychometrics in disciplines such as geology – to analyze compositional survey data. Aitchison suggested to log transform ratios of compositional responses, turning them into convenient differences, which Coenders and colleagues then modelled using a multitrait-multimethod (MTMM) design. Embedding compositional data in the SEM framework is a very attractive proposition in psychometrics, since latent variables of substantive interest – such as personality traits or attitudes – may be measured. Unlike in survey data, where the focus is on estimating parameters of stimuli (such as population means or covariances of each alternative in a composition), the focus in psychometric tests is on estimating person parameters (for example, person score on Extraversion). To date, however, no model has been suggested to infer proper measurement of individual differences from multiplicative ipsative questionnaire data. This article aims to address this gap.

The article is organized as follows. First, the compositional data analysis tradition based on the seminal work of Aitchison (1982) is applied to responses collected within personality and similar questionnaires. It is shown that in the context of psychological assessment, the units of analysis – the log-transformed ratios of points – are readily interpretable as the difference of utilities that respondents feel for questionnaire items. This

interpretation enables the use of log-ratios as the continuous observed outcomes in an SEM framework. Thurstonian factor models are used to model the mean and covariance structure of the log-ratios. The article develops identification constraints and other technical detail required for estimating these models, and discusses how to deal with zeros that may be present in the compositions. To prove that the method recovers well the true item parameters and the true latent trait correlations, a simulation study is conducted using a simple compositional design and two sample sizes. Finally, an empirical data analysis example is provided to illustrate the approach.

Compositional Questionnaire Data Analysis

Compositional Format and Response Process

Consider blocks consisting of $n \geq 2$ stimuli (here, questionnaire items), among which respondents have to distribute a fixed number of points C (for example, $C = 100$) according to some instruction, for instance, the extent to which the items describe respondents' personality, or reflect their attitudes, etc. Regardless of the exact values of n and C, the points assigned to each item divided by the block total are proportions – hence such blocks are called *compositions*, and collections of such blocks *compositional data* (Aitchison, 1982). Here is an example block, in which a hypothetical respondent distributed 100 points according to the extent the adjectives described his/her personality:

	Points
A. Dependable	50
B. Curious	20
C. Modest	20
D. Calm	10

Since all points add to a constant, questionnaires in this format give rise to ipsative data. Regardless of the absolute psychological values a respondent may attach to the items, his/her responses reveal only relative strengths of preferences within blocks.

We can presume that the observed composition is a result of a response process, in which respondent *j* evaluates the actual psychological values (v_{j1}, v_{j2}, …, v_{jn}) he/she feels for the items, but is able to express them only as proportions of the given total *C*,

$$y_{ji} = C v_{ji} \Big/ \sum_{q=1}^{n} v_{jq} \, . \tag{1}$$

We may also assume the values v_{ji} on a ratio scale, with 0 representing no value to the respondent, and the ratio *x* between two items meaning that the first item has *x* times the value to the respondent compared to the second item.

Transformation of Compositions into Differences of Utilities

The compositional format constrains respondents to express only proportions of the psychological values they feel for the items; however, the responses maintain the original ratios between the values:

$$\frac{y_{ji}}{y_{jk}} = \frac{C v_{ji} \Big/ \sum_{q=1}^{n} v_{jq}}{C v_{jk} \Big/ \sum_{q=1}^{n} v_{jq}} = \frac{v_{ji}}{v_{jk}} . \tag{2}$$

From the responses in our earlier example, we may infer that: (1) the psychological value of A to the respondent was five times greater than the value of D (ratio A/D = 5); (2) the value of B was two times greater than D (ratio B/D = 2); (3) the value of C was two times greater than D (ratio C/D = 2), etc. Note that the ratios of three items (arbitrarily, A, B and C) to the remaining item (arbitrarily, D) capture information about the composition fully. From these ratios, one can derive that the psychological values of B and C were equal (B/D:C/D =

B/C = 1); or that the value of A was two and a half times stronger than the values of either B or C (A/B = A/C = 2.5).

More generally, responses of person *j* to a compositional block consisting on *n* items can be fully described by *n* – 1 ratios of points, whereby ratios of all but one items to a referent item *k* (arbitrarily, the last item in the block) are computed. Each ratio y_{ji}/y_{jk} reflects how many times the value of item *i* is greater (or smaller) than the value of item *k* for the respondent.

INSERT FIGURE 1 ABOUT HERE

Figure 1 (panel a) illustrates a typical distribution of ratios of points given to questionnaire items within the same composition. It can be seen that the distribution is approximately lognormal. It was Aitchison's (1982) idea to transform the bounded-by-zero and positively skewed ratios of compositional data using the natural logarithm function,

$$y_{j\{i,k\}} = \ln\left(\frac{y_{ji}}{y_{jk}}\right) = \ln\left(y_{ji}\right) - \ln\left(y_{jk}\right), \tag{3}$$

to yield outcome variables $y_{j\{i,k\}}$ that are unbounded and approximately normally distributed (see Figure 1, panel b). The advantage of the log-ratio transformation is that it places compositional data in the unconstrained "multivariate real space, opening up all available standard multivariate techniques" (Aitchison & Egozcue, 2005; p. 831). Indeed, the transformation converts unworkable ratios into convenient differences. Given that the ratios of original psychological values are preserved in the observed data, as equality (2) shows, the log-ratios of observed scores y_{ji} represent the differences of logarithms of the latent psychological values v_{ji}. We can label the logarithms of latent values, $\ln(v_{ji})$, as t_{ji}

$$y_{j\{i,k\}} = \ln\left(v_{ji}\right) - \ln\left(v_{jk}\right) = t_{ji} - t_{jk}\,. \tag{4}$$

After log-ratio transformations have been applied to the observed responses, the resulting pairwise outcomes $y_{j\{i,k\}}$ can be conceptualized as the differences of arbitrarily scaled item *utilities* t_{ji}. Utility is a well-established concept introduced by Thurstone to describe the "affect that the object calls forth" (Thurstone, 1929; p.160). Here, we use the term for two reasons. First, to separate the "utility" t_{ji} from the previously used "value" v_{ji}, which, although representing the same psychological phenomenon are scaled and distributed differently. While ratio-scaled values v_{ji} are distributed log-normally, interval-scaled utilities t_{ji} are distributed normally. Second, to connect to the large body of literature on comparative data analysis using Thurstonian law of comparative judgment (Thurstone, 1927), to which the notion of utility is central. Indeed, very clear parallels can be drawn between choice behavior driven by utility maximization, and assigning values in compositions. Thus, when the ratio of observed points in (3) is greater than one, the pairwise outcome $y_{i\{i,k\}}$ is positive, indicating that the first item in the pair has higher utility than the second item. In a choice task, this utility judgment would result is selection of the first item over the second. When the ratio of observed points in (3) is less than one, the pairwise outcome $y_{i\{i,k\}}$ is negative, indicating that the first item in the pair has lower utility than the second item. When the ratio of points is exactly 1, the pairwise outcome is zero, indicating that the two items in the pair have equal utilities.

Treatment of Zeros in Compositions

A challenge to computing log-ratios arises whenever respondents reject one or more items completely, assigning those zero points. Ratios including zeros yield either zero (when item i is given 0 points and item k is given a positive number of points) or infinity (when item k is given 0 points), for which natural logarithms cannot be computed. An effective solution to this problem was given by Martín-Fernández, Barceló-Vidal and Pawlowsky-Glahn (2003), who suggested replacing any zero with a fixed imputed value δ, which is smaller than

the smallest number of points actually possible to express in compositions ("smallest detectable value"). In geology applications where compositional analysis was originally developed, the rationale for this replacement strategy is that zeros typically result from rounding in measurement of very small values, or insufficient sensitivity of a measurement tool. In psychological assessment, it is reasonable to assume that zero may be a result of a very small subjective psychological value that falls below the smallest integer that can be provided as a response (or, more precisely, below a threshold that separates the felt value from the smallest integer). For example, if the smallest positive number of points that can be given to any one alternative is 1, then any value that feels to a respondent subjectively "smaller" than that gets expressed as 0.

Because the replacement of zeros with imputed values δ distorts the original compositions, non-zero responses also have to be adjusted to preserve the total C and the ratios among responses. The following replacement formula

$$y_{ji(r)} = \begin{cases} \delta, & \text{if } y_{ji} = 0 \\ y_{ji}\left(1 - \frac{1}{C}\sum_{y_{ji}=0}\delta\right), & \text{if } y_{ji} \neq 0 \end{cases} \quad (5)$$

is recommended since it preserves the compositions (Martín-Fernández et. al., 2003). Specifically, ratios for all non-zero values are preserved, which also ensures preservation of the covariance structure of non-zero elements of compositions. The latter feature is extremely important to psychometric applications considered in the present paper.

When deciding on the actual value of δ to use for imputation, it is reasonable to aim for a value that is representative of a typical uncensored latent response. Research with geological compositional data showed that when the proportion of zeros in data was below 10%, the replacement procedure performed best with $\delta = 0.65$ of the threshold (or the smallest detectable value; Martín-Fernández et al., 2003). Sandford, Pierson and Crovelli

(1993) suggested imputing value $\delta = 0.55$ of the threshold in the same type of applications. However, in psychological applications the actual threshold for assigning either 0 or the smallest permissible number of points (e.g. 1) is not known. A sensitivity analysis on the choice of the imputed value is recommended in such applications, particularly if the amount of zeros is substantial.

Thurstonian Factor Models for Differences of Utilities

The main goal of analysis of questionnaire data is to model broader factors underlying item responses (personal attributes that a questionnaire is designed to measure). Conceptualizing the observed variables in compositional questionnaires as the differences of item utilities provides a straightforward connection to established models of choice data – Thurstonian factor models (Maydeu-Olivares & Böckenholt, 2005). These models have been applied to forced-choice questionnaires (Brown & Maydeu-Olivares, 2011a; 2012), where the observed comparative judgments are binary (prefer item *i* or item *k*). In such choice formats, the differences of utilities are not observed, only their dichotomizations are observed. In the case of compositional data, the differences of utilities are observed directly. These are the pairwise log-ratios (4) – continuous variables, which we assume normally distributed. This section describes Thurstonian factor models for continuous outcomes, and shows how to estimate these models.

Mean and covariance structure of utility differences. The utility judgment about a questionnaire item is assumed a random process, with systematic influences from psychological attribute(s) the item is designed to measure and a random error. The most common model for item utilities is a linear factor model (e.g., McDonald, 1999),

$$t_{ji} = \mu_i + \sum_{a=1}^{d} \lambda_{ia} \eta_{ja} + \varepsilon_{ji}, \tag{6}$$

where μ_i is the mean of utility t_i, $\eta_{j1}, \eta_{j2}, ..., \eta_{jd}$ are factor scores (latent traits) weighted by factor loadings $\lambda_{i1}, \lambda_{i2}, ..., \lambda_{id}$, and ε_{ji} is the unique factor. Because respondents are sampled randomly from the population of interest, we can treat all person-specific parameters (subscripted *j*) as random effects, and present the utilities (6) in matrix form

$$\mathbf{t} = \boldsymbol{\mu} + \boldsymbol{\Lambda}\boldsymbol{\eta} + \boldsymbol{\varepsilon}. \quad (7)$$

The common factors $\boldsymbol{\eta} = (\eta_1, \eta_2, ..., \eta_d)'$ are distributed across people as multivariate normal with the covariance matrix $\boldsymbol{\Phi}$. The error terms $\boldsymbol{\varepsilon}$ are normally distributed and uncorrelated with the common factors and with each other so that their covariance matrix $\boldsymbol{\Psi}^2$ is diagonal. In a questionnaire with *p* compositional blocks containing *n* items each, there are *pn* items, therefore the vector of item means $\boldsymbol{\mu}$ contains *pn* elements. $\boldsymbol{\Lambda}$ is a ($pn \times d$) matrix of the factor loadings of *pn* items on *d* factors. Most often, questionnaire items are designed to measure one factor only, so that the matrix of factor loadings $\boldsymbol{\Lambda}$ has only one non-zero entry in every row (has an *independent clusters basis*; McDonald, 1999). The items within one block may indicate the same trait (unidimensional comparisons) or different traits (multidimensional comparisons).

According to expression (4), the observed variables and the units of analysis in compositional questionnaires – the log-ratios of items *i* and *k* – are the differences of item utilities, $y_{j\{i,k\}} = t_{ji} - t_{jk}$. Because each composition of *n* items yield only $n - 1$ pairwise outcome variables , with *p* compositions there are $p(n - 1)$ observed variables, written in matrix form as

$$\mathbf{y} = \mathbf{A}\mathbf{t}. \quad (8)$$

In this expression, $\mathbf{A}$ is a ($p(n - 1) \times pn$) block-diagonal design matrix, representing the contrasts between all but one item in a block to the referent item (arbitrarily, the last item).

When $n = 2$, each block in $\mathbf{A}$ is $\mathbf{A_2} = \begin{pmatrix} 1 & -1 \end{pmatrix}$, whereas when $n = 3$, and $n = 4$, they are, respectively,

$$\mathbf{A}_3 = \begin{pmatrix} 1 & 0 & -1 \\ 0 & 1 & -1 \end{pmatrix}, \qquad \mathbf{A}_4 = \begin{pmatrix} 1 & 0 & 0 & -1 \\ 0 & 1 & 0 & -1 \\ 0 & 0 & 1 & -1 \end{pmatrix}. \tag{9}$$

Expressions (8) and (7) give the basis of modeling the observed log-ratios as a higher-order factor structure. Figure 2 illustrates this structure for a hypothetical test measuring four latent traits with three blocks of size $n = 4$. It can be seen that the latent utilities $\mathbf{t}$ are modelled as first-order factors according to (8). Each observed log-ratio is determined by two latent utilities (note that there is no error term in equation (8)); the utility loadings are fixed to 1 and −1 respectively. All but one of n utilities in each block is indicated by one observed log-ratio; only the utility of the referent item in the block is indicated by all n–1 log-ratios. The latent traits $\boldsymbol{\eta}$ are modelled as second-order factors according to (7). The second-order factors are indicated by the latent utilities of their respective items; the factor loadings are freely estimated. The first-order latent variables – the utilities – have error/disturbance terms $\boldsymbol{\varepsilon}$.

INSERT FIGURE 2 ABOUT HERE

Given the assumption of normally distributed log-ratios, only the means and covariances are needed to describe the observed data. The mean and covariance structure of $\mathbf{y}$ is given by

$$\boldsymbol{\mu}_y = \mathbf{A}\boldsymbol{\mu} = \boldsymbol{\gamma}, \text{ and } \quad \boldsymbol{\Sigma}_y = \mathbf{A}\left(\boldsymbol{\Lambda}\boldsymbol{\Phi}\boldsymbol{\Lambda}' + \boldsymbol{\Psi}^2\right)\mathbf{A}', \tag{10}$$

where $\boldsymbol{\gamma}$ is the ($p(n-1)$) vector of intercepts replacing the differences of utility means[i].

Estimable parameters. The following fixed parameters are estimated in a questionnaire with p compositions, each containing n items:

1. **Intercepts**. One intercept $\gamma_{\{i,k\}}$ is estimated for each outcome variable $y_{\{i,k\}}$, making $p(n-1)$ estimable intercepts in total.
2. **Factor loadings** (the elements of $\boldsymbol{\Lambda}$). These are the factor loadings of item utilities. When every item measures only one attribute, one loading per item is estimated, pn loadings in total. Items may measure more than one attributes; more factor loading parameters are estimated in this case.
3. **Error variances**, the diagonal elements of $\boldsymbol{\Psi}^2$**.** These are the residual variances of item utilities, pn in total.
4. **Factor covariances**, the matrix $\boldsymbol{\Phi}$. There are $d(d-1)/2$ covariances to estimate.

Model identification. To identify the model, one needs to set metrics for the second-order factors $\boldsymbol{\eta}$, the first-order factors $\mathbf{t}$, and the residuals $\boldsymbol{\varepsilon}$. The second-order factors' (traits) variances are set to one, and their means are set to zero. The first-order factors' (utilities) means and the means of their residuals are set to zero. Conveniently, all latent variable means are set to zero by default in Mplus. These are the only identification constraints needed in the general case[ii].

A special case arises when compositions contain only two items ($n = 2$). In this case, the residual variances of the two utilities underlying just one observed outcome are not separately identified, and need to be constrained equal to identify the model. An additional special case arises when compositions consist of two items ($n = 2$), each measuring one of two assessed attributes ($d = 2$). Because this model is essentially an exploratory factor model (i.e. each observed log-ratio variable indicates both second-order factors), additional identification constraints need to be imposed on some factor loadings.

The important feature of measurement of individual differences with compositional questionnaires, and other comparative response formats (e.g. forced-choice format), is that the scales of the latent traits are generally identified (i.e. their proper covariance matrix $\mathbf{\Phi}$ can be estimated). Thus, ipsative data do not arise, and interpersonal comparisons can be made. Brown (2014) shows that identifiability of the latent traits is made possible by multiple items indicating each trait, and a full-rank matrix of contrast loadings $\mathbf{A\Lambda}$. Empirical non-identification is possible when, for example, factor loadings within every block are equal, or when factor loadings within every attribute are equal. This is in contrast to the general indeterminacy of the scale origin of the item utilities (Böckenholt, 2004), which results from impossibility to uniquely determine n utilities from $n - 1$ contrasts. For instance, one cannot estimate the covariances of utilities from single compositions without imposing further constraints (Maydeu-Olivares & Böckenholt, 2005), which is a natural limitation in applications focused on stimuli rather than broader traits underlying them.

Model and person parameters estimation. The mean and covariance structure (10) of the utility differences can be estimated using general-purpose SEM software, using maximum likelihood estimation. Mplus (L.K. Muthén & B.O. Muthén, 1998-2012) conveniently combines all necessary features[iii]. After the model parameters have been estimated, person scores on the attributes measured by a questionnaire may be estimated. When all outcomes are continuous, the regression method with correlated factors (Lawley & Maxwell, 1971) is used for estimating the factor scores by Mplus. Conveniently, Standard Errors for all estimated traits are also provided.

Simulation Study: Estimating Item Parameters in a Short Compositional Test

To investigate how well the proposed approach can recover true model parameters, a simulation study was carried out. The study considered an extremely simplified

compositional test with four traits measured by three blocks of four items (quads), or 12 items in total. The structural model for the test is presented in Figure 2; an Mplus input code to test this model is provided in Appendix A. Each trait is measured by exactly three items; the first item in each block measures Trait 1, the second item in each block measures Trait 2, etc. The true model parameters for this design were taken from the numerical Example 2 in Brown and Maydeu-Olivares (2012), to enable parallelism and easy comparison between the present study using the compositional format and the previously published study using the forced-choice format. The true generating parameter matrices – the item intercepts $\boldsymbol{\mu}$, the factor loadings $\boldsymbol{\Lambda}$, the residual variances $\boldsymbol{\Psi}^2$, and the trait covariances $\boldsymbol{\Phi}$ – are provided in Appendix B.

The study generated 12 normally distributed utilities conforming to the given factor structure (with the given intercepts, factor loadings and residual variances), and the given covariances between the latent traits. Note that this stage does not involve compositional data at all – the generated utilities can be directly analyzed with the usual confirmatory factor model for rating scale data. Next, the differences of utilities were computed in each block between the three first items and the last (referent) item according to (4). The compositional model depicted in Figure 2 was then fitted to the utility differences. The model estimated 39 parameters (12 factor loadings, 12 error variances, 9 intercepts and 6 factor correlations), and had 15 degrees of freedom.

Two sample sizes were tested in the simulation study, N = 300 and N = 1000, with 1000 replications each. To determine the percentage of *parameter bias*, the true parameter was subtracted from the average parameter value across 1000 replications; the result was divided by the true parameter and multiplied by 100. To determine *standard error bias*, the standard deviation of the parameter estimates across 1000 replications (with the large number

of replications this is considered *population standard error*) was subtracted from the average of the estimated standard errors across replications; the result was divided by the population standard error and multiplied by 100.

The results were very encouraging for both the smaller and the larger sample sizes. For N = 300, the average χ^2 was 15.745 (df = 15), with the rejection rate .073, which was only slightly higher than the nominal rate .05. The parameter bias ranged from –3.65% to 1.40%, and the standard error bias ranged from –3.95% to 5.68%. For N = 1000, the average χ^2 was 15.119 (df = 15), with the rejection rate .047, extremely close to the nominal rate .05. The parameter bias ranged from –1.29% to 0.56%, and the standard error bias ranged from –5.96% to 6.03%. For both sample sizes, the utility residual variances were least accurately estimated, with predominantly negative parameter bias (the residual variances were slightly underestimated). The other parameters were estimated very precisely, with the average bias close to zero.

Empirical Study: A Big Five Measure in Compositional Format

Materials

Items from the English version of the Forced-Choice Five Factor Markers (FCFFM; Brown & Maydeu-Olivares, 2011b) were used in this study. The items measure broad markers of the Five Factors of personality (Neuroticism, Extraversion, Openness to experience, Agreeableness and Conscientiousness). The FCFFM questionnaire consists of 60 behavioral statements (e.g. “I leave a mess in my room”), with 12 statements measuring each of the Five Factors. The 60 items are organized in 20 blocks of three (triplets) so that each item in a block measures different trait, and equal numbers of pairwise comparisons are made between different traits. For more detail on the rationale and development of the FCFFM see Brown and Maydeu-Olivares (2011b).

In the original version of the questionnaire, respondents have to rank order the statements within blocks according to the extent the statements are true of them. For the purpose of this study, the response format was changed so that the blocks of three items became compositions, in which respondents had to distribute $C = 15$ points according to the extent the statements were true of them. In addition, the respondents were asked to rate the statements using a 5-point scale ("very untrue" – "somewhat untrue" – "in between" – "somewhat true" – "very true"). Compositional and rating tasks for the same three items were performed consecutively: first the respondents rated block 1, then provided compositional ratings for block 1, then moved to block 2, etc.

Participants and Procedure

Psychology students from a UK university completed the questionnaire online in return for research credits. Out of N = 317 participants, 80.1% were female. Age ranged from 18 to 51 years (median = 19.0; mean = 19.9; SD = 4.3 years). The participants were asked to "complete a short personality questionnaire using a conventional rating format and an alternative format", and were then debriefed.

Analyses

Sensitivity analysis to determine the optimal imputed value. Before fitting a CFA model to the compositional questionnaire responses, any zeros had to be imputed with a small value δ. To determine the optimal imputed value in the present questionnaire, sensitivity analyses were conducted. Typically, such analyses would measure discrepancies between "true" data (data with no zeros but potentially very small values present) and imputed data (where any small values below the "smallest detectable value" or threshold would be replaced with δ), for multiple chosen values δ. The imputed value yielding the smallest discrepancy with the true data would be then selected as optimal. In the present study,

however, the "true" psychological values behind "censored" zero entries are not known. An acceptable strategy in this situation would be to use available proxies of the items' psychological values v_{ji}, from which "true" compositions can be inferred using (1). Fortunately, Likert ratings of the items can help obtain such proxies.

The Likert ratings can be assumed the observed indicators for the underlying item utilities t_{ji}. Given the relationship between the utility and the psychological value $t_{ji} = \ln(v_{ji})$, the psychological values were obtained as $v_{ji} = \exp(t_{ji})$, and then were transformed into 20 compositions of three items using (1). The resulting compositions were considered proxies for "true" compositions[iv], which possessed the necessary feature of absence of zeros (i.e. all resulting values were positive). In the constrained simplex space, the discrepancy between the proxy "true" composition $\mathbf{v}_j = (v_{j1}, v_{j2}, \ldots v_{jn})$ and the observed composition $\mathbf{y}_j = (y_{j1}, y_{j2}, \ldots y_{jn})$ of n items for person j is the Aitchison distance (Aitchison, 2002)

$$d_A\left(\mathbf{v}_j, \mathbf{y}_j\right) = \left(\sum_{i=1}^{n} \left(\ln\left(y_{ji} / g\left(\mathbf{v}_j\right)\right) - \ln\left(v_{ji} / \mathrm{g}\left(\mathbf{y}_j\right)\right)\right)^2 \right)^{1/2}. \tag{11}$$

The Aitchison distance is simply the Euclidean distance between two log-ratio transformations centred at the geometric mean of each composition, $g\left(\mathbf{x}_j\right) = \left(\prod_{i=1}^{n} x_{ji} \right)^{1/n}$.

With this, the overall measure of discrepancy for each block across N respondents is the mean squared Aitchison distance,

$$\mathrm{msd} = \frac{1}{N} \sum_{j=1}^{N} d_A^2\left(\mathbf{v}_j, \mathbf{y}_j\right). \tag{12}$$

The value δ that minimized the mean squared distances across all 20 compositions was considered optimal, and was adopted for imputing any zero values in observed responses.

Fitting the measurement models. First, the rating scale (single-stimulus) responses were analyzed using a straightforward confirmatory factor model with five correlated traits.

The five-point ratings were treated as continuous data, which is considered a reasonable approach to analysis of ordinal responses with five or more categories (Rhemtulla, Brosseau-Liard, & Savalei, 2012). Twelve observed item scores indicated each of the five factors.

Second, the compositional responses to the FCFFM were analyzed. After replacing all zero responses with the optimal value δ based on the result of the sensitivity analysis, and adjusting the remaining items in the compositions using formula (5), 40 log-ratios were computed, two in each of the 20 blocks, with the last item in each block used as referent. A higher-order factor model was fitted, with 60 latent utility variables underlying the 40 observed log-ratios, and five second-order factors (the Big Five) underlying the latent utilities. Each second-order factor was indicated by 12 latent utilities.

Both the rating scale and compositional models were fitted in Mplus 7.2, using robust maximum likelihood estimator. To judge goodness of fit, we considered the chi-square statistic (χ^2), the Standardized Root Mean Square Residual (SRMR), and the Root Mean Square Error of Approximation (RMSEA). When testing covariance structures, SRMR values less than .08 are thought to indicate good fit; for RMSEA, values less than .06 are thought to indicate good fit (Hu & Bentler, 1999).

Person scores and their standard errors. For each response format, model-based factor scores were estimated for each participant and saved by Mplus. The variances of the estimated factor scores and their Standard Errors (SE) are also printed. The SE values were squared to obtain the estimated population error variances of each scale. Reliability of each scale was computed using the classic definition – as the proportion of variance in the estimated factor score $\hat{\eta}$ due to the true score η:

$$\rho = \frac{\mathrm{var}(\hat{\eta})}{\mathrm{var}(\eta)} = \frac{\mathrm{var}(\eta) - SE^2(\eta)}{\mathrm{var}(\eta)}. \quad (13)$$

Results

Sensitivity analysis to determine the optimal imputed value. The 20 compositions varied in the proportions of zero values present. The proportions ranged from 1.1% zeros among its 3*317 observed responses for the composition {i13, i14, i15} to 10.7% zeros for the composition {i25, i26, i27}. The average proportion of zeros among all the compositions was 4.7%. Ten δ values between 0.1 and 0.9 with an increment 0.1 were tried for imputation, for which the mean square Aitchison distances (msd) to the corresponding 20 compositions derived from ratings were computed. For any given value δ, the msd values for the 20 compositions were distributed with a large positive skew. Therefore, the median msd across the 20 compositions was judged the best measure of central tendency. The median msd for all imputed values δ are plotted in Figure 3 (the mean msd are also plotted). It can be seen that the median discrepancy between the imputed compositional data and the rating scale proxies is the largest for the smallest value δ = .1, it then rapidly decreases, reaches a minimum at δ = .5, and then slowly increases again as the δ value approaches the maximum of 0.9. The same is true for the mean msd. Exactly the same shape of the msd function when "true" small values were known was observed by Martín-Fernández et al. (2003). As the result of this sensitivity analysis, we chose to use the common value **δ = .5** for imputing all zeros in the empirical example.

INSERT FIGURE 3 ABOUT HERE

Measurement model for rating scale data. The exact fit of the model for rating scale responses was relatively poor, with $\chi^2 = 3897.13$ on 1700 degrees of freedom ($p < 0.001$), SRMR = .086; although approximate fit was almost acceptable with RMSEA = .064 (90 percent confidence interval .061-.066). The largest modification index ($\chi^2 = 126.08$)

pertained to correlated residuals of item 25 (“I love to read challenging material”) and item 36 (“I avoid difficult reading material”). The next largest modification index ($\chi^2 = 67.72$) was for correlated residuals of items 29 (“I get irritated easily”) and 47 (“I rarely get irritated”); and the next ($\chi^2 = 62.59$) was for correlated residuals of item 49 (“I leave a mess in my room”) and item 56 (“I like to tidy up”). The remaining modification indices were much lower in magnitude ($\chi^2 = 30$ or less). As can be seen, all areas of misfit can be easily understood as shared specific item content within the respective broad personality factors, rather than any cross-loadings or other problems.

After adding correlated residuals for the three pairs of items identified above, the modified model fitted slightly better, with $\chi^2 = 3602.65$ on 1697 degrees of freedom ($p < 0.001$), SRMR = .085, RMSEA = .060 (90 percent confidence interval .057-.062). The model-based correlations of the five latent traits are given in Table 1 (above the diagonal).

INSERT TABLE 1 ABOUT HERE

Measurement model for compositional data. The compositional model yielded better goodness of fit than the rating scale model. Specifically, it had reasonable exact fit, $\chi^2 = 1209.88$ on 690 degrees of freedom ($p < 0.001$), SRMR = .072; and good approximate fit, RMSEA = .049 (90 percent confidence interval .044-.053). Just like in the rating model, the largest modification index ($\chi^2 = 38.42$) pertained to correlated residuals of items 25 and 36. The next largest was the index $\chi^2 = 31.06$ pertaining to correlated residuals of item 18 (“I often forget to put things back in their proper place”) and item 49 (“I leave a mess in my room”). Other modification indices were of magnitude $\chi^2 = 17$ or less. As in the rating model, the areas of misfit were due to similarity of item content within their respective scales, which was over and above their shared variance due to the broad personality factor.

After adding correlated residuals for the two pairs of item identified above (that is, the residuals of the first-order latent utilities were correlated), the modified model fitted slightly better, with $\chi^2 = 1135.96$ on 688 degrees of freedom ($p < 0.001$), SRMR = .070, RMSEA = .045 (90 percent confidence interval .041-.050). The estimated correlations between the five latent dimensions in the compositional model are given in Table 1 (below the diagonal). It can be seen that the correlations yielded by both rating scale and compositional models were largely similar.

Score reliability. The standard errors and reliabilities of the estimated scale scores in the rating scale and compositional models are given in Table 2. All scales in both format yielded reliable scores in the range of .8–.9, except the scale Openness, which yielded low reliability in the compositional format ($\rho = .611$) but not in the rating scale format ($\rho = .812$). For other scales, the scores were slightly more reliable when the Likert ratings were used.

Relationships between rating scale and compositional factor scores. Estimated factor scores from the two CFA models were used to explore the relationships between corresponding scales (hetero-method mono-trait correlations), which are given in Table 2. The estimated trait scores of the same concepts correlated highly, and were similar in magnitude to their respective reliability coefficients. The correlation coefficients corrected for unreliability of both estimated scores are provided in Table1 (on the diagonal). It can be seen that except the trait Agreeableness, for which the corrected correlation was .914, the other traits' corrected correlations were very close to 1.

INSERT TABLE 2 ABOUT HERE

Conclusions

The present article extends the Thurstonian modeling approach beyond binary choice data to quantitative preferences collected in the form of compositions, or the "proportion-of-total" response format. Analysis of compositional data has an established tradition based on the seminal work of Aitchison (1982). Here, this tradition is adopted for conceptualizing compositional responses collected within personality and similar questionnaires as proportional reflection of the strengths of psychological values that respondents feel for questionnaire items. Despite the constraint on the total of all the psychological values within each composition, the ratios of observed points preserve the ratios of the unobserved psychological values. This enables the use of ratios of points awarded to items within compositions to infer the psychological values, and through the values to infer the psychological attributes the questionnaire is designed to measure. All information contained in compositional blocks of size n is fully described by $\tilde{n} = n - 1$ ratios of points to an arbitrarily chosen referent item k (for example, the last item in the block). These pairwise ratio variables are distributed approximately log-normally; the log transformation is applied to the ratios to achieve approximately normally distributed outcome variables.

The log-ratios of item points are the units of compositional analysis; they can be thought to represent arbitrarily scaled pairwise differences of items' utilities (Thurstone, 1927). The mean and covariance structure of log-ratios is analyzed using confirmatory factor analysis, assuming that two latent utilities determine each observed log-ratio, and that a number of second-order factors (attributes the questionnaire measures) underlie the latent utilities. The model is estimated using maximum likelihood, and the person attribute scores are estimated by the regression method with correlated factors (Lawley & Maxwell, 1971).

The effectiveness of the proposed estimation procedure was assessed in a simulation study, where normally distributed utilities conforming to a simple factor structure were

generated. The differences of generated utilities representing arbitrarily scaled ratios of points in a hypothetical compositional questionnaire with blocks of size $n = 4$ were analyzed using the proposed approach with a second-order CFA model. The estimated parameters were very close to the true parameters; and the rejection rates were very close to nominal for both, a smaller sample with N = 300 and a larger sample with N = 1000. The conclusion from this study is that estimation of compositional questionnaire models is robust with the minimum sample size usually recommended for CFA (N = 300).

Following recommendations of Martín-Fernández et al. (2003), any zeros in item scores are replaced with a small non-zero value before computing the log-ratios, to avoid zero and infinity ratios. After this non-parametric imputation procedure, the compositions are adjusted to maintain the original ratios of non-imputed values. This is important for ensuring that the covariance structure of the compositions stays intact. The imputed value must be smaller than the smallest non-zero value possible to submit as a response. Although general recommendations can be made (e.g. replacing zeros with .5 of the smallest admissible response); the exact choice may be determined by sensitivity analyses in each particular application. Considering the relative complication to otherwise straightforward analysis imposed by the presence of zeros in compositions, it may be sensible to bar participants from entering zero values, particularly when the total number of points C is large. This can be easily achieved in computerized administrations, where validation on the permissible entries can be implemented easily.

Finally, the proposed approach was illustrated with empirical data. A comparison of confirmatory factor analyses based on compositional and Likert-type responses to the Forced-Choice Five Factor Markers (FCFFM) using a sample of N=317 students was carried out. Sensitivity analyses, which minimized the Aitchison distances between the imputed compositions and the Likert-based proxies for item utilities, yielded the optimal imputed

value $\delta = .5$. A second-order CFA model was fitted to the imputed compositions, yielding a reasonable goodness of fit. The hypothesized response process and factorial structure were confirmed and cross-validated by very close similarity with the constructs based on Likert ratings. It is concluded that the proposed approach is a straightforward and effective way of analyzing compositional questionnaire data.

Endnotes

[i] Utility means are usually not of interest and are not estimated here. There are more means to estimate in each block (n) than there are observed intercepts ($n - 1$); thus the means sub-model is over-parameterized. The means can be estimated only when additional constraints are implemented; for instance, the mean of one referent item in each block can be set to 0. For more detail, see Maydeu-Olivares & Böckenholt (2005).

[ii] For the reader familiar with Thurstonian IRT modelling, the identification constraints needed for compositional data are exactly the same as imposed with binary forced-choice data, except there is no need to set the residual variance of one utility per block, since these are identified with continuous outcomes.

[iii] An Excel macro, which automates syntax building for testing compositional questionnaire data using Mplus can be obtained from the author upon request.

[iv] Considering the compositions of ratings rather than the ratings themselves gives the advantage of removing any uniform response biases that may be present in ratings, making the rating data more robust proxies for true item utilities.

References

Aitchison, J. (1982). The statistical analysis of compositional data. *Journal of the Royal Statistical Society. Series B (Methodological), 44* (*2*), 139-177.

Aitchison, J., & Egozcue, J. J. (2005). Compositional data analysis: Where are we and where should we be heading? *Mathematical Geology, 37(7)*, 829-850.

Böckenholt, U. (2004). Comparative judgments as an alternative to ratings: Identifying the scale origin. *Psychological Methods, 9*, 453–465.

Brown, A. (2014, December 10). Item Response Models for Forced-Choice Questionnaires: A Common Framework. *Psychometrika*. Advance online publication. doi: 10.1007/s11336-014-9434-9

Brown, A. & Maydeu-Olivares, A. (2011a). Item response modeling of forced-choice questionnaires. *Educational and Psychological Measurement, 71*, 460-502.

Brown, A. & Maydeu-Olivares, A. (2011b). *Forced-Choice Five Factor Markers*. Retrieved from PsycTESTS. doi: 10.1037/t05430-000

Brown, A. & Maydeu-Olivares, A. (2012). Fitting a Thurstonian IRT model to forced-choice data using Mplus. *Behavior Research Methods*, *44*: 1135–1147.

Chan, W. (2003). Analyzing ipsative data in psychological research. *Behaviormetrika, 30*, 99–121.

Cheung, M.W.L, & Chan, W. (2002). Reducing uniform response bias with ipsative measurement in multiple-group confirmatory factor analysis. *Structural Equation Modeling*, *9*, 55-77.

Clemans, W. V. (1966). An analytical and empirical examination of some properties of ipsative measures. *Psychometric Monographs*, 14.

Coenders, G., Hlebec, V., & Kogovsek, T. (2011). Measurement Quality in Indicators of Compositions. A Compositional Multitrait-Multimethod Approach. *Survey Research Methods, 5(2)*, 63-74.

Hu, L. T., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling, 6(1)*, 1-55.

Kahneman, D. (2011). *Thinking, fast and slow*. London, UK: Allen Lane.

Lawley, D. N., & Maxwell, A. E. (1971). *Factor analysis as a statistical method*. London: Butterworths.

Martín-Fernández, J.A., Barceló-Vidal. C., & Pawlowsky-Glahn, V. (2003). Dealing with zeros and missing values in compositional data sets using non-parametric imputation. *Mathematical Geology, 35*, 253–278.

Maydeu-Olivares, A. *&* Böckenholt, U. (2005). Structural equation modeling of paired-comparison and ranking data. *Psychological Methods*, *10*, 285-304.

Maydeu-Olivares, A. & Brown, A. (2010). Item response modeling of paired comparison and ranking data. *Multivariate Behavioral Research, 45*, 935 - 974.

McDonald, R. P. (1999). Test theory: A unified treatment. Mahwah, NJ: Erlbaum.

Muthén, L.K. *&* Muthén, B.O. (1998-2012). *Mplus User's guide. Seventh edition*. Los Angeles, CA: Muthén & Muthén.

Rhemtulla, M., Brosseau-Liard, P. É., & Savalei, V. (2012). When can categorical variables be treated as continuous? A comparison of robust continuous and categorical SEM estimation methods under suboptimal conditions. *Psychological Methods, 17*, 354-373.

Sandford, R. F., Pierson, C. T., & Crovelli, R. A. (1993). An objective replacement method for censored geochemical data. *Mathematical Geology, 25 (1)*, p. 59–80.

Thurstone, L.L. (1927). A law of comparative judgment. *Psychological Review*, *34, 273-286*.

Thurstone, L.L. (1929). The measurement of psychological value. In Smith and Wright (eds), *Essays in Philosophy by Seventeen Doctors of Philosophy of the University of Chicago*. Chicago: Open Court, 157-174.

Appendix A. Mplus Syntax for Fitting the Model Presented in Figure 2

```
DATA: FILE = 3quads.dat;

VARIABLE:
        NAMES =y1-y12;  !observed compositions
        USEVARIABLES   !these are the log-ratios produced by the DEFINE command
        i1i4 i2i4 i3i4
        i5i8 i6i8 i7i8
        i9i12 i10i12 i11i12;

DEFINE:
        ! computation of log-ratios for 3 blocks
        i1i4=ln(y1/y4);   i2i4=ln(y2/y4);   i3i4=ln(y3/y4);
        i5i8=ln(y5/y8);   i6i8=ln(y6/y8);   i7i8=ln(y7/y8);
        i9i12=ln(y9/y12);  i10i12=ln(y10/y12);   i11i12=ln(y11/y12);

ANALYSIS: ESTIMATOR=MLR;

MODEL:
        !utilities - first order factors
        t1  BY  i1i4@1;         t2  BY  i2i4@1;         t3  BY  i3i4@1;
        t4  BY  i1i4@-1 i2i4@-1 i3i4@-1;

        t5  BY  i5i8@1;         t6  BY  i6i8@1;         t7  BY  i7i8@1;
        t8  BY  i5i8@-1 i6i8@-1 i7i8@-1;

        t9  BY  i9i12@1;        t10  BY  i10i12@1;    t11  BY  i11i12@1;
        t12  BY  i9i12@-1 i10i12@-1 i11i12@-1;

        !errors of log-ratios are zero since they are determined by the utility differences
        i1i4-i11i12@0;

        !latent traits - second order factors
        F1 BY t1* t5 t9;
        F2 BY t2* t6 t10;
        F3 BY t3* t7 t11;
        F4 BY t4* t8 t12;

        F1-F4@1;

OUTPUT: STDY;   !standardized solution

SAVE:  !saves estimated trait scores and their SEs
        FILE=3quadsResults.dat; SAVE=FSCORES;
```

Appendix B. Population Parameters for the Simulation Study

$$\boldsymbol{\mu} = \left(0, 0.5, -1, 0.5, 0.5, 0.2, 0.2, -0.3, -0.5, 1, 1.5, 0\right)'$$

$$\boldsymbol{\Psi}^2 = diag\left(1,1,1,1,1,1,1,1,1,1,1,1\right)$$

$$\boldsymbol{\Phi} = \begin{pmatrix} 1 & -0.4 & 0 & 0.4 \\ -0.4 & 1 & 0.3 & -0.3 \\ 0 & 0.3 & 1 & 0 \\ 0.4 & -0.3 & 0 & 1 \end{pmatrix}$$

$$\boldsymbol{\Lambda} = \left(\begin{array}{cccc} 1 & 0 & 0 & 0 \\ 0 & -0.8 & 0 & 0 \\ 0 & 0 & 1.3 & 0 \\ 0 & 0 & 0 & 0.8 \\ \hline -1.3 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 0.8 & 0 \\ 0 & 0 & 0 & 1.3 \\ \hline 0.8 & 0 & 0 & 0 \\ 0 & 1.3 & 0 & 0 \\ 0 & 0 & -1 & 0 \\ 0 & 0 & 0 & 1 \end{array}\right)$$

Table 1

Estimated correlations between the Big Five markers based on the rating scale and compositional versions of the FCFFM in the empirical example

	N	E	O	A	C
Neuroticism (N)	1.015	-.258**	-.193**	.033	.132
Extraversion (E)	-.303**	1.012	.307**	.195**	-.091
Openness (O)	-.184*	.139	1.070	.295**	.030
Agreeableness (A)	.107	.115	.350**	.914	.171*
Conscientiousness (C)	.228**	-.207*	-.249*	.046	.983

Note: The mono-method hetero-trait latent correlations from the *rating scale* model are **above** the diagonal, from the *compositional* model are **below** the diagonal. The hetero-method mono-trait correlations of estimated factor scores corrected for unreliability are on the diagonal. ** Correlations are significant at the .01 level, two-tailed. * Correlations significant at the .05 level, two-tailed.

Table 2

Standard errors of measurement and reliabilities of the Five Factor markers based on the rating scale and compositional versions of the FCFFM in the empirical example

	Rating scale			Compositional			
	$SE(\eta)$	$\text{var}(\eta)$	ρ	$SE(\eta)$	$\text{var}(\eta)$	ρ	$\text{corr}(\eta_R, \eta_C)$
Neuroticism (N)	.306	.907	.897	.367	.865	.844	.883**
Extraversion (E)	.298	.911	.903	.383	.853	.828	.875**
Openness (O)	.398	.841	.812	.529	.720	.611	.754**
Agreeableness (A)	.319	.898	.887	.399	.841	.811	.775**
Conscientiousness (C)	.339	.885	.870	.423	.821	.782	.811**

Note: ** Correlations are significant at the .01 level, two-tailed.

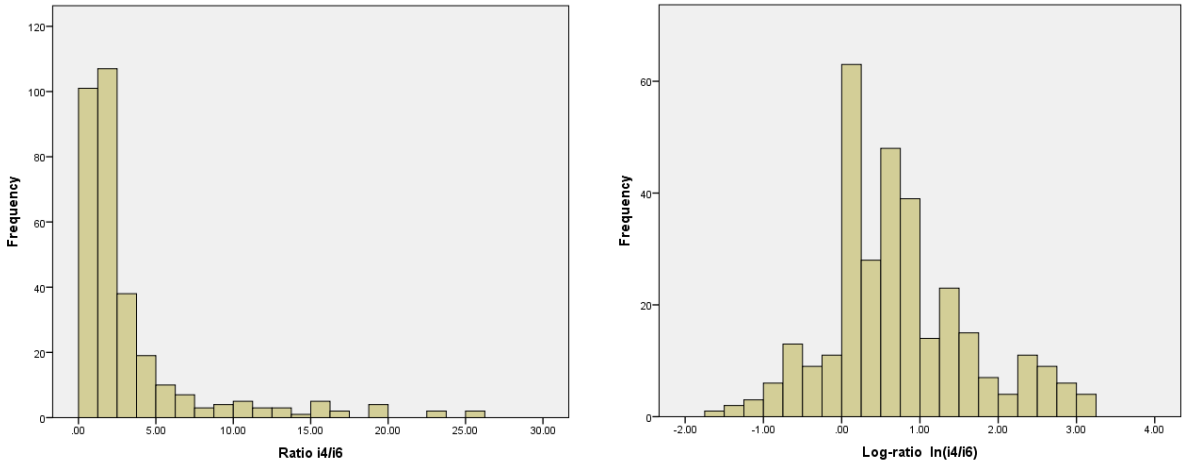

Figure 1. Example distribution of ratios of points given to two questionnaire items within the same composition, and the corresponding log-ratio.

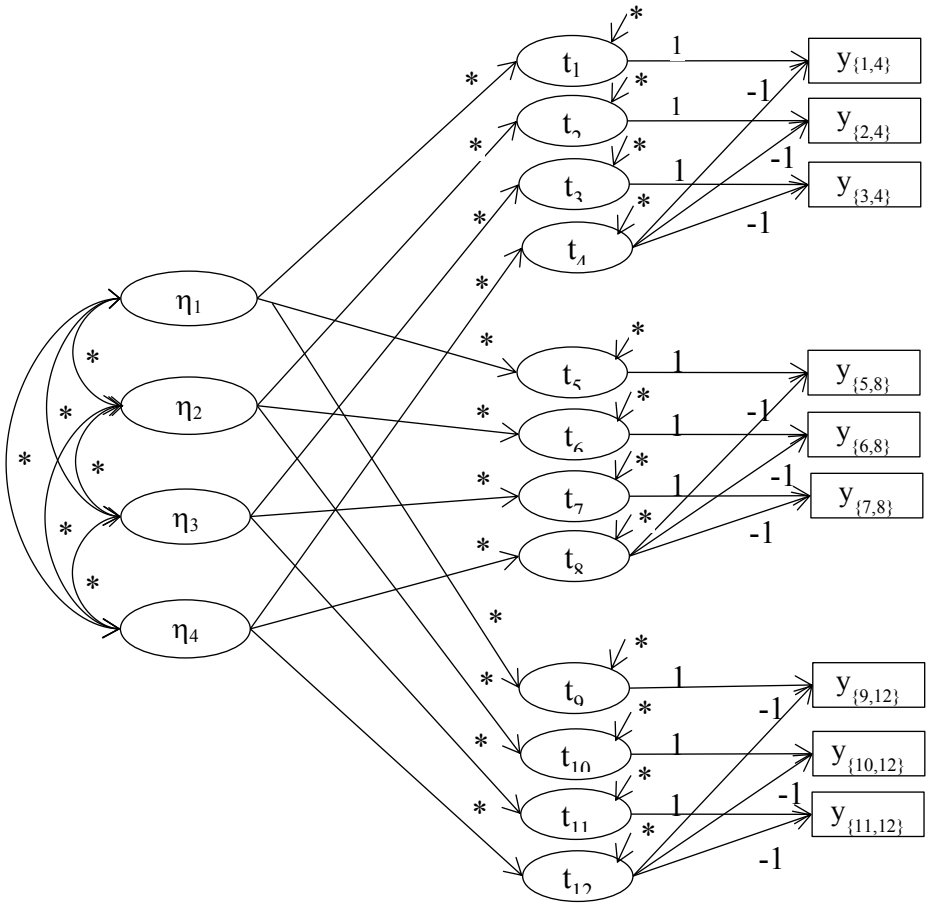

Figure 2. Measurement model for a simple design consisting of three compositional blocks of four items (quads) used in the simulation study.

Note. Asterisks mark freely estimable parameters. There are 30 parameters pertaining to the covariance structure shown in this Figure, plus there are 9 intercept parameters (one per each observed outcome) that are not shown. To scale the second-order factors η, their variances are set to 1 and means to 0; to scale the utility errors, their means are set to 0.

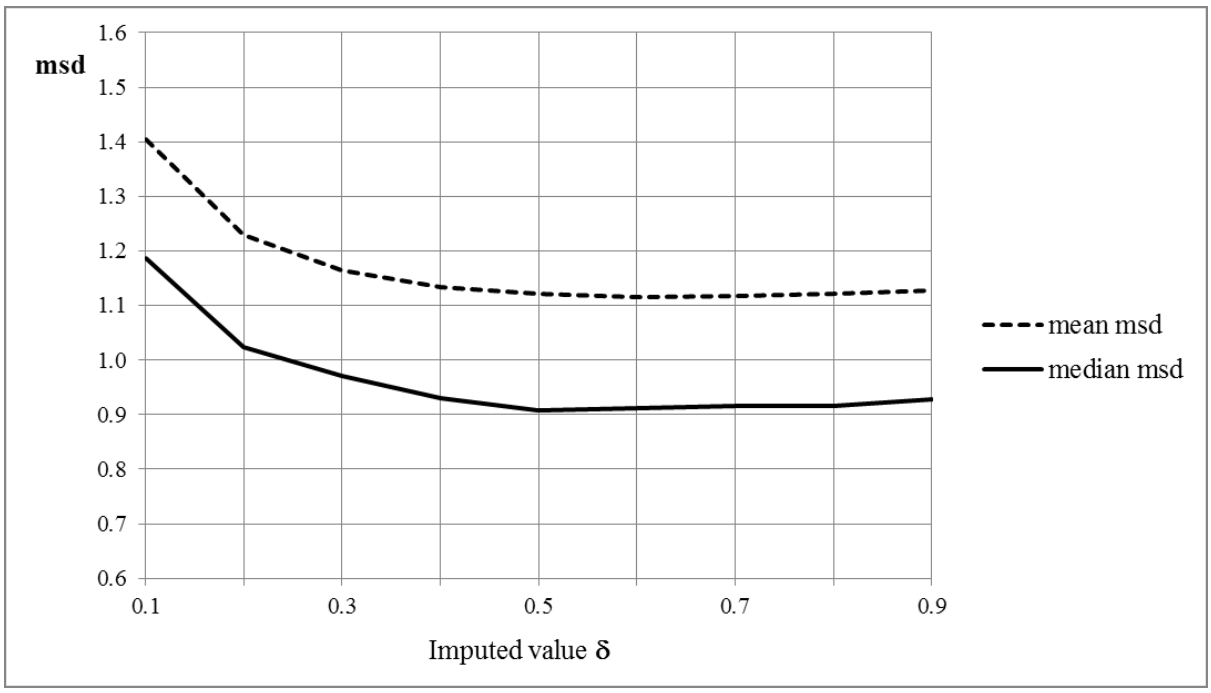

Figure 3. Sensitivity of measure of distortion (the mean and the median of "msd" between imputed data and compositions of proxy values across 20 blocks) to changes in the imputed value δ